**בס"ד**

**בהנחיית: פרופ' אוריאל פייגה  
מגישה: אילנה ברזניק  
אוקטובר 2012**

תוכן עניינים:

מבוא..............................................................................................3

מהו גרף..........................................................................................4

הצגת גרף בעזרת מטריצה.................................................................5

אינטרפרטציה גרפית למושג הערך העצמי במטריצות...........................8

רקע תיאורטי Spectral Graph Theory.............................................9

מציאת ערך עצמי מקסימלי של מטריצה בעזרת ה Power Method....16

הרעיון של קליינברג........................................................................18

שאילתות ומקורות מוסמכים.............................................................19

ניתוח הרשת המתוקשרת...................................................................19

בניית גרף מרוכז באינטרנט................................................................20

חישוב רכזים ודפים איכותיים............................................................22

האלגוריתם האיטראטיבי..................................................................23

רשימה ביבליוגרפית.......................................................................26

מבוא

מטרתי בעבודה זו היא להציג את אחת השיטות מהשנים האחרונות לדרוג הדפים באינטרנט. השיטה הוצגה על ידי קליינברג במאמרו JON M. KLEINBERG/Authoritative Sources in a Hyperlinked Environment/*Journal of the ACM, Vol. 46, No. 5, September 1999, pp. 604-632*. טענתו המרכזית של קליינברג היא כי ניתן ללמוד הרבה על איכות הדפים של הסביבה המתוקשרת בעזרת ניתוח המבנה שלה. תנאי הכרחי ללמידה זו הוא הימצאותם של הכלים המתאימים. בעבודתו המוצגת במאמר מציג קליינברג כלים אלגוריתמיים שמטרתם לנתח את מבנה האינטרנט ואת הקשרים בו ככלי לקבלת אינפורמציה על איכות הדפים בו. באמצע שנות ה-90, מנוע החיפוש של Yahoo דרג את הדפים באינטרנט בצורה ידנית. כיום, האינטרנט כבר גדול מידי ולכן שיטה זו הופכת להיות יקרה ולא אפקטיבית. השאלה הגדולה היא כיצד לעשות את הדרוג בצורה אוטומטית?. אפשרות אחת היא להמציא תוכנה הקוראת את דפי האינטרנט ומדרגת אותם לפי איכותם. כיום תוכנה כזו טרם הומצאה עקב המורכבות שבעניין. לא נפסלת האפשרות כי תוכנה כזו תומצא בעתיד. אפשרות אחרת היא לחשוב על מדד שייתן אינדיקציה על איכות הדפים באינטרנט ותוך כדי שהדבר יוכל להיעשות בצורה אוטומטית על ידי מחשב. עבודתו של קליינברג היא מהראשונות המציעות איך לעשות את דרוג הדפים לא רק לפי בדיקת תוכנם אלא על ידי בחינת דפי האינטרנט המצביעים עליהם, שהם מהווים כמעין "ממליצים". לצורך בדיקת איכותו של דף אינטרנט לא מספיק לבדוק את כמות הממליצים. שכן, אם זהו הקריטריון לבחינת חשיבותו של דף האינטנרט ניתן לבנות בצורה מלאכותית הרבה דפי אינטרנט שיצביעו על דף זה. קליינברג מציע קריטריון יותר מורכב. בנוסף להתייחסות לכמות הממליצים, הוא רוצה להתייחס גם לאיכותם. האלגוריתם שמציע קליינברג מתקשר לחישוב ערכים עצמיים של מטריצה המייצגת את הקשרים בין דפי האינטרנט. בעבודה זו, נסקור את האלגוריתם של קליינברג ואת הרקע התיאורטי הדרוש להבנתו.

מהו גרף

גרף הינו ייצוג מופשט של [קבוצה](http://he.wikipedia.org/wiki/%D7%A7%D7%91%D7%95%D7%A6%D7%94_(%D7%9E%D7%AA%D7%9E%D7%98%D7%99%D7%A7%D7%94)) של אובייקטים, כאשר כל תת-קבוצה של אובייקטים בקבוצה עשויים להיות מקושרים זה לזה. האובייקטים הניתנים לקישור מכונים קודקודים או צמתים (באנגלית: vertex), וקבוצת הקודקודים מסומנת באות \ V. הקישורים בין הקודקודים מכונים צלעות או קשתות (באנגלית: edge), וקבוצת הצלעות מסומנת באות \ E. מתקיים כי קבוצת הצלעות מקיימת: E \sube V\times V, כלומר: כל צלע הינה זוג הקודקודים, אותם היא מקשרת. גרף, אשר קבוצת הקודקודים שלו היא \ Vוקבוצת הצלעות שלו היא \ Eמסומן באופן הבא: \ G=(V,E).

לדוגמא, נתבונן בגרף הבא:





**תורת הגרפים הספקטלרית** היא תורה המאפשרת להתאים לכל גרף מטריצה וללמוד מהערכים העצמיים של המטריצה את תכונות הגרף.

הצגת גרף בעזרת מטריצה

ניתן לייצג גרף בדוגמא על ידי כמה מטריצות:

**1. מטריצת קודקודים-קשתות:**



1

2

3

4

מטריצה זו מוגדרת על הקודקודים והקשתות על ידי ההגדרה הבאה:



אם קודקוד j עובר דרך צלע i

אחרת

**2. מטריצת הסמיכויות:**

מוגדרת על הקודקודים.



אם סמוך ל

אחרת

בדוגמא שלנו המטריצה היא:

4 3 2 1

1



2

3

4

**3. מטריצת הלפלסיאן:**

מטריצה זו מוגדרת על הקודקודים בלבד.



אם



אם וגם סמוך ל



אחרת

הצורה המנורמלת של מטריצה זו היא :



אם וגם



אם וגם סמוך ל

אחרת

כאשר הגרף הוא רגולרי, כלומר הדרגות של כל הקודקודים שוות, שתי המטריצות שקולות.

נבנה את המטריצות עבור הגרף שבדוגמא.

המטריצה הרגילה:

4 3 2 1

1



2

3

4

המטריצה המנורמלת:

4 3 2 1



1

2

3

4

נשים לב שבגרפים לא מכוונים המטריצות הנ"ל הן סימטריות.

צורה נוספת להסתכל על מטריצת הלפלסיאן היא להסתכל על מטריצה המתקבלת מההפרש בין מטריצת הדרגות למטריצת הסמיכויות של גרף G. על מטריצת הדרגות היא מטריצה סקלרית שבה מופיעות על האלכסון דרגות הקודקודים.

לדוגמא, בגרף הקודם:



סימון:



הערכים העצמיים והוקטורים העצמיים של מטריצות אלו מספקים אינפורמציה על הגרף. עבור גרפים רגולריים, המידע שניתן להסיק מאחת ההצגות של המטריצה זהה למידע שניתן להסיק מההצגות האחרות. לעומת זאת, בגרפים שאינם רגולריים, בחירת ההצגה היא דבר המשנה מאוד.

הסבר לכך שבגרף רגולרי הייצוג לא משנה: ניתן לעשות טרנספורמציה פשוטה בין הערכים העצמיים של מטריצת הסמיכויות לערכים העצמיים של מטריצת הלפלסיאן.

במקרה של גרף d רגולרי הטרנספורמציה היא:



מטריצת הסמיכויות

הע"ע מוכפל ב 1 - נוסף d לע"ע

הו"ע לא משתנה הו"ע לא משתנה

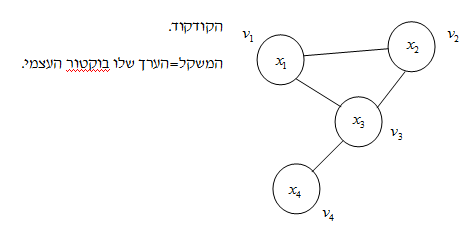
הערה: לאורך כל העבודה נשתמש במשפט מאלגברה לינארית שאם המטריצה היא ממשית וסימטרית אז הערכים העצמיים שלה ממשיים, והווקטורים העצמיים מהווים בסיס ל . כלומר הם פורסים ובלתי תלויים.

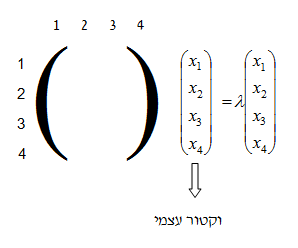


אינטרפרטציה גרפית למושג הערך העצמי במטריצות

אנו רגילים מאלגברה לינארית לאינטרפרטציה אלגברית למושג הערך העצמי.

עבור מטריצה ,  הוא וקטור עצמי עם ערך עצמי , אם מתקיים .

נסתכל על אינטרפרטציה נוספת למושג הערך העצמי, שהיא אינטרפרטציה גרפית.





סכום המשקלים של הקודקוד .



לדוגמא: מסכמים את המשקל של השכנים (המחוברים אליו בקשת) של ומקבלים .

רקע תיאורטי Spectral Graph Theory

משפט 1: 0 הוא תמיד ערך עצמי של מטריצת הלפלסיאן .

הוכחה: (1,1,...1,1) הוא תמיד וקטור עצמי כי כל רכיב בוקטור הוא למעשה סכום האיברים בכל שורה במטריצה . הדבר קורה מכיוון ש  הוא וקטור האחדים. סכום זה הוא בדיוק 0 כי הוא מורכב מ:

הדרגה של כל קודקוד+(1-)\*מס' הקודקודים הסמוכים לו=0.

משפט 2: מספר הפעמים שבהם 0 הוא ערך עצמי של מטריצת הלפלסיאן הוא כמספר רכיבי הקשירות של הגרף.

הוכחה: נניח שלגרף G , K רכיבי קשירות בגודל . נחלק את המטריצה לפי רכיבי קשירות. נסדר את המטריצה כך ש  השורות הראשונות של L יהיו שייכות לרכיב הראשון.  השורות הבאות יהיו שייכות לרכיב השני וכך הלאה.... נקבל מטריצה מהצורה:



L=

לפי משפט באלגברה לינארית: אם מחליפים את סדר השורות ובאותו האופן את סדר העמודות הערכים העצמיים לא משתנים כי פשוט שינינו את סדר הקודקודים במטריצה. המטריצה  היא תת-מטריצה  המייצגת את מטריצת הלפלסיאן של הרכיב ה-i של .G נגדיר את הוקטור  כוקטור עם קואורדינטות 1 בכל הקואורדינאטות המתאימות לרכיב ה-i ו 0 בשאר הקואורדינאטות. נעשה זאת עבור i=1,…,k .



כל הוקטורים הם בת"ל כי המכפלה הפנימית של כל שניים מהם היא 0 מה שאומר שהם ניצבים. לכן יש ריבוי לערך העצמי כי יש לערך עצמי זה כמה וקטורים בלתי תלויים.



נסתכל על וקטור מהצורה: .

ניתן לראות כי : Ly=0y.

אינטואיציה: התת גרף המיוצג על ידי  הוא בעל רכיב קשירות אחד ולכן הוקטור (1,1,...1,1) הוא וקטור עצמי שלו עם ערך עצמי 0.

התרומה של שאר הרכיבים  היא 0 כי הגדרנו 0 בשאר הקואורדינאטות. לכן קיבלנו כי (0,...0, 1, ..., 1) הוא וקטור עצמי עם ערך עצמי 0.



גודל רכיב הקשירות הראשון.

באותו האופן, ניתן להגדיר עוד K-1 וקטורים עצמיים נוספים, על ידי שניקח 1 ברכיבים של אותו רכיב קשירות ו 0 בשאר. בצורה זו נקבל K בחירות בלתי תלויות של הוקטור y . ולכן קיבלנו בסיס לתת-מרחב המתאים לערך העצמי 0 של מטריצת הלפלסיאן.

משפט 3:במטריצת הסמיכויות הערך העצמי המקסימלי בגרף d רגולרי הוא d.

הוכחה: גרף d רגולרי הוא גרף שבו לכל קודקוד יש d שכנים. כלומר בכל שורה של מטריצת הסמיכויות יש d אחדים. מכיוון שהמטריצה סימטרית גם בכל עמודה יש d אחדים. נניח  הוא הערך העצמי מקסימלי.

**הכיוון :** נסתכל על וקטור האחדים. מכאן עם ערך עצמי ולכן הערך עצמי המקסימלי מקיים . **הכיוון :**  נניח  וקטור עצמי ונניח . נרצה להראות שהערך העצמי קטן או שווה d. נפעיל על הוקטור  את המטריצה A. . מכיוון שהמטריצה A היא מטריצת הסמיכויות, משמעות הפעולה היא כי מסכמים את משקלי הקודקודים הסמוכים ל  וזה יהיה הערך של  וכו'. נניח .  כלומר  ולכן . אי שיווין זה יהיה נכון לכל קואורדינאטה i ולכן הערך העצמי יהיה קטן או שווה ל d. משני כיוונים אלו נובע כי הערך העצמי המקסימלי הוא בדיוק d.

משפט 4:במטריצת הסמיכויות הע"ע המקסימלי:א. לא יותר גדול מהדרגה המקסימלית.

ב. לא יותר קטן מהדרגה הממוצעת.

הוכחת 4 א': גרף לא רגולרי הוא גרף שבו לא לכל קודקודים יש את אותה הדרגה.

נרצה להוכיח  כאשר  הוא הערך העצמי המקסימלי. נניח הוא וקטור עצמי ונניח , . נרצה להראות שהערך העצמי קטן או שווה ל . נפעיל על הוקטור y את המטריצה A. Ay=z. מכיוון שהמטריצה A היא מטריצת הסמיכויות, משמעות הפעולה היא כי מסכמים את משקלי הקודקודים הסמוכים ל וזה יהיה  וכו'... נניח, .  כלומר  ולכן .

אי שוויון זה נכון לכל קואורדינאטה  ולכן הערך העצמי יהיה קטן או שווה . הוכחת 4 ב':נסמן את הדרגה הממוצעת על ידי  כאשר  היא דרגת הקודקודים. נרצה להוכיח כי  כאשר. נוכיח זאת בשני חלקים שמהם תנבע הטענה.

**חלק 1':** . הטענה נכונה לכל וקטור יחידה.

**חלק 2':** נרצה למצוא וקטור יחידה  שעבורו מתקיים .



**הוכחת חלק 1'**: נניח כי הם הוקטורים העצמיים ו  הם הערכים העצמיים מתאימים להם. מהווים בסיס ל ונניח שכל הוקטורים העצמיים הם מנורמלים כלומר .

נסתכל על הביטוי  ונבדוק מהו ערכו עבור וקטור יחידה . ניתן להציג את כצירוף לינארי של אברי הבסיס : . ולכן לפי משפט פיתגורס ה-n מימדי .

הסבר: הוקטורים העצמיים ניצבים כי  עבור . ולכן קיבלנו כי .

**הוכחת חלק 2':**ניקח את  להיות וקטור האחדים המנורמל. כלומר .

A היא מטריצת הסמיכויות. לכן מספר האחדים בה הוא כסכום דרגות הקודקודים.

נחשב את ערך הביטוי .



 =סכום דרגות הדוקדוקים

הסבר:

 כאשר  היא דרגת הקודקוד ה  בגרף.



סכום דרגת הקודקודים

הערה: הערך הכי גדול של הביטוי הוא הערך העצמי המקסימלי וזה קורה כאשר נשים במקום y את הוקטור העצמי המתאים לו.

משפט5: בגרף דו-צדדי: גרף שקודקודיו מחולקים לשתי קבוצות זרות U,V כך שכל קשת מחברת קודקוד מ-U עם קודקוד מ-V, הערכים העצמיים באים בזוגות. כלומר אם  ערך עצמי אז גם  ערך עצמי. הוכחה: נניח קודקודי הקבוצה U הם: . וקודקודי הקבוצה V הם: . טענה: אם  וקטור עצמי עם ערך עצמי 

אז  וקטור עצמי עם ערך עצמי .

הוכחה: נראה שהדבר מתקיים עבור נציג מהקבוצה U,  ועבור נציג מהקבוצה V, .

מההנחה,  כאשר מחובר בקשת ל .

ולכן גם  כאשר מחובר בקשת ל .

כל משקלי הקודקודים מצד ימין של השוויון שייכים לקודקודים מהקבוצה V ולכן משקלם לא משתנה.

בנוסף,  כאשר מחובר בקשת ל .

ולכן גם  כאשר מחובר בקשת ל . כל משקלי הקודקודים מצד ימין של השוויון שייכים לקודקודים מהקבוצה U ולכן משקלם מחליף סימן.

משפט 6: הערכים העצמיים של מטריצת הלפלסאן הם גדולים שווים ל 0.

הלפלסיאן היא מטריצה מוגדרת חיובית Positive-definite matrix)). כלומר סימטרית וללא ערכים עצמיים שליליים.

הוכחה: בגרף רגולרי קל לראות זאת כי:

מטריצת הסמיכויות



במטריצת הסמיכויות אין ערך עצמי הגדול מ d ולכן במטריצת הלפלסיאן אין ערכים עצמיים שליליים.

בגרף לא רגולרי,

הראינו כי  כאשר  הוא הערך העצמי המקסימלי ו  הוא הערך העצמי המינימלי.

נראה כי לכל וקטור ,  כאשר  היא מטריצת הלפלסיאן. מכאן ינבע כי כל הערכים העצמיים גדולים או שווים ל 0.

ניזכר כי מטריצת הלפלסיאן היא מהצורה:

כאשר  היא דרגת הקודקוד ה .



בכל מקום אחר, יש 1- אם ישנה קשת המחברת בין הקודקודים ו 0 אחרת.



ניתן לפרק את המטריצה לסכום של תת מטריצות המייצגות את הקשתות.

K i j i



K j

מייצגת את הקשת (i,k) מייצגת את הקשת (j,k)

הערה: בשאר הקואורדינאטות יש 0.

נסמן:

= התת מטריצה המתאימה לקשת המחברת בין i ל j.

= התת מטריצה המתאימה לקשת המחברת בין i ל k.

ואז כאשר ישנה קשת המחברת בין n ל m.



לפי חוק הפילוג, התוצאה תהיה שווה לסכום ביטויים מהצורה:

j i



i

j

מכיוון שהסכום מורכב ממחוברים שהם גדולים או שווים ל 0 הוא גם כן גדול או שווה ל 0.

מכאן, למטריצת הלפלסיאן אין ערכים עצמיים שליליים.

משפט 7: היא מטריצה סימטרית כאשר , .

הוכחה: ניזכר בהגדרת כפל מטריצות:

נניח, , , מטריצות , .

אזי:



נרצה שיתקיים: כאשר: . נראה שזה אכן מתקיים.



ומכאן ולכן היא מטריצה סימטרית כנדרש.

באותה צורה ניתן להראות כי היא מטריצה סימטרית.

מציאת ערך עצמי מקסימלי של מטריצה בעזרת ה **Power Method**:

**שיטה 1:** בעזרת ליכסון. מציאת הפולינום האופייני והשורשים שלו. בעזרת שיטה זו מוצאים את כל הערכים העצמיים. **שיטה 2:** **(The Power Method).** אם מעוניינים רק בערך העצמי המקסימלי, ישנה צורה יותר מהירה מהשיטה הקודמת. לא מובטח כי השיטה תעבוד בוודאות, אבל היא תעבוד בהסתברות גבוהה. כלומר זה עובד באופן מעשי. בעזרת שיטה זו מוצאים גם את הוקטור העצמי . נראה כיצד היא עובדת:

נרצה . ניקח וקטור שרירותי כלשהו . נפעיל עליו את המטריצה ונקבל: . כעת, על מנת להימנע מערכים גדולים בקואורדינאטות, ננרמל את הוקטור .



נמשיך הלאה ונקבל: 



 כאשר .

נניח הוקטורים העצמיים הם:  ונניח שהם מסודרים בסדר עולה. מכיוון שהם מהווים בסיס למרחב הוקטורי ,

כל וקטור ניתן להצגה כצרוף לינארי שלהם. בפרט, הוקטור .

 ולכן:





 נמשיך כך ונקבל:

 מתכנס לוקטור המתאים לערך העצמי הכי גדול אם נחזור על התהליך מספיק פעמים.

אם  הכי גדול אז כל שאר הרכיבים נהיים זניחים. מכאן מוצאים את הוקטור העצמי.

ככל שחוזרים על התהליך יותר פעמים, מתקרבים יותר לוקטור העצמי.

אחרי שמצאנו את הוקטור העצמי נציב אותו במשוואה ההתחלתית ונמצא את הערך העצמי. . הערות:

• כאשר בוחרים באקראי את הוקטור , חשוב שנבחר אותו כך שלא יהיה ניצב לוקטור העצמי המתאים לערך העצמי הגדול ביותר . מכיוון ש , אם  ניצב ל  אז . לכן אם המקדם של הערך העצמי הגדול ביותר הוא אפס הוא מתבטל וההתכנסות תהיה לערך העצמי הבא בתור מבחינת גודל. הסיכוי שנעשה בחירה לא מתאימה של הוקטור  הוא מאוד קטן.

• כמה מהר נתכנס לוקטור העצמי? השאלה היא מהם שאר הערכים העצמיים. אם הערכים העצמיים השייכים לאותם וקטור עצמי הם רחוקים, ההתכנסות תהיה מהירה. ואם לא, זה ייקח יותר זמן.

•בעיה: בגרף דו-צדדי. בין הערכים העצמיים של מטריצת הסמיכויות מופיעים בזוגות .

זו אכן בעיה כי אם נעלה בחזקה זוגית, הערך המוחלט שלהם יהיה אותו הדבר.

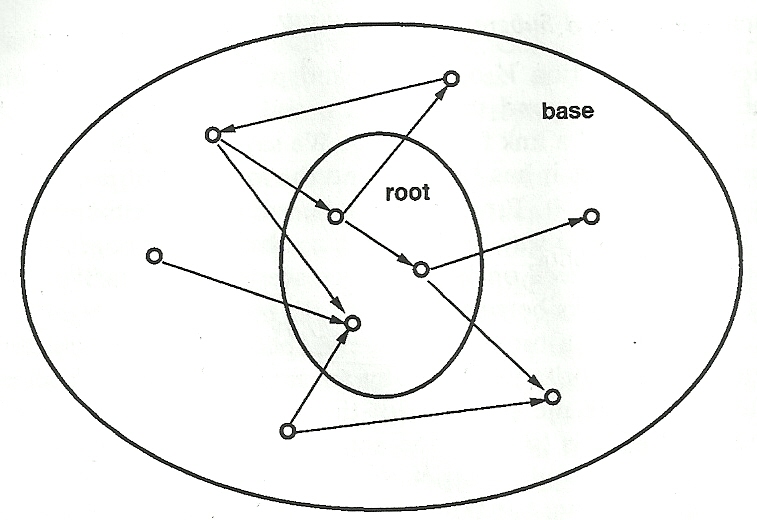
הרעיון של קליינברג

הרעיון של קליינברג הוא להחשיב כדפים ממליצים רק את הדפים שהם איכותיים בעצמם. אך לשם כך שוב נשאלת השאלה מהו דף ממליץ איכותי?. אחת האפשרויות היא להחשיב כממליץ איכותי דף ששאר הדפים שהוא ממליץ עליהם הם איכותיים גם כן. למעשה נכנסנו לתהליך הבנוי משלבים שחוזרים על עצמם כאשר בכל שלב עלינו לבדוק את איכות דף האינטרנט. תהליך זה נקרא תהליך איטראטיבי (Iterative ). תהליך שבו חוזרים על אותה פעולה שוב ושוב כאשר כל פעולה בודדת נקראת איטראציה. לכן התהליך שקליינברג מציע הוא תהליך איטראטיבי. קליינברג בעבודתו מתייחס לשני סוגי דפים. Authority = דף איכותי שנקרא לו דף סמכות. דף בעל איכות טובה עם תוכן מעניין על פי נושא חיפוש מסויים. Hub =דף רכז. דף שממנו קל להגיע לדפים איכותיים, הוא מהווה ריכוז של הרבה הצבעות כלומר הוא מצביע על הרבה אתרים טובים. לכן הרכזים עוזרים לגלות דפים איכותיים. לדוגמא, אם אנו מחפשים דפים בעלי איכות טובה עם אינפורמציה על כלבים אז דף רכז יכול להיות פורום של מגדלי כלבים. דף רכז הוא מעין צומת מרכזית שכדאי לעבור דרכה על מנת להגיע למקומות חשובים. מפאת גודלה של הרשת, לא היינו רוצים להפעיל את אלגוריתם הדרוג על כל הרשת כולה. לכן בדרך כלל יש שלב איתחול הבוחר חלק מהרשת. במאמר לא נאמר בדיוק איך עושים זאת אך בדרך כלל שלב זה מורכב משני חלקים. החלק הראשון הוא סינון הדפים שהכילו את מילת החיפוש בעצמה. החלק השני הוא הרחבת הקבוצה זו על ידי הוספת הדפים שמצביעים עליהם דפים מהחלק הראשון. נרצה לדמות את חלק האינטרנט שבחרנו לגרף דו-צדדי, גרף שקודקודיו מחולקים לשתי קבוצות, דפים איכותיים ודפי רכז. כאשר ישנן קישורים בין שתי קבוצות של דפים אלו אך לא בתוך הקבוצות עצמן. ניתן להסתכל על הדפים כעל קודקודים ועל הקישורים כעל קשתות מכוונות. האינטואיציה לרעיון של קליינברג היא כי אם מופיע באינטרנט מבנה של גרף דו-צדדי צפוף מעין זה, כנראה שמסתתרים כאן דפים בעלי תוכן איכותי. לצורך אישוש ההשערה אפשר פשוט להסתכל מדגמית על כמה מהם ולראות אם אלו דפים איכותיים או לא. במקרה וחברה לקידום אתרים מגלה את השיטה היא יכולה לבנות מבנה מסוג זה סביב דף אינטרנט שברצונה לקדם. לכל תת גרף כזה נוכל להתאים מטריצה הנקראת מטריצת הסמיכויות. בעזרת הווקטורים העצמיים של מטריצה זו נוכל ליצור דרוג לקבוצת הדפים המיוצגים על ידי הגרף. המאמר מתמודד עם השאלה איך בכלל צריך לגשת לעניין דרוג הדפים. עדיין לא מדובר על איך להפוך את האלגוריתם ליעיל, אלא קודם צריך להבין היטב את השאלה ואת הדרישות מהתשובה בצורה ברורה.

**שאילתות ומקורות מוסמכים** שאילתה היא המשפט שהמשתמש כותב בשורת מנוע החיפוש. ישנם שלשה סוגי שאילתות. סוג א': נושא חיפוש ספציפי. זו שאלה ספציפית לדוגמא: מהו מזג האוויר מחר בירושלים? סוג ב': נושא חיפוש רחב. לדוגמא: מצא מידע אודות התכנות בשפת ג'אווה. סוג ג': בקשה למציאת דפים דומים. לדוגמא: מצא דפים דומים ל java.sun.com. נתמקד בשני הסוגים הראשונים. בכל אחד מהם, הקשיים בחיפוש הם שונים. בחיפוש ספציפי, יש מעט מאוד דפים שיכילו את מילות המפתח המבוקשות ולכן ייתכן שיהיה קשה למצוא את האינפורמציה המבוקשת. לעומת זאת, בחיפוש רחב יהיו אלפים רבים של דפים שיתאימו להגדרת החיפוש. לכן על מנת לאפשר חיפוש אפקטיבי, חייבת להימצא דרך לסנן את הדפים המוסמכים ביותר מתוך המגוון הרב הקיים. המושג "דף איכותי" קשור לשאילתת נושא החיפוש הרחב. מושג זה יהווה את מקור ההתמקדות המרכזי שלנו במאמר. המכשול המרכזי בעניין הוא כיצד נקבע אם דף אינטרנט מסויים הוא דף איכותי עבור שאילתה מסויימת או לא?

**ניתוח הרשת המתוקשרת** ניתוח הרשת המתוקשרת יעזור לנו להתמודד עם הקשיים שהעלנו עד כה. למעשה, הקישורים באינטרנט טומנים בתוכם שיפוט אנושי סמוי והטענה שלנו היא שזהו בדיוק השיפוט אותו אנו צריכים כדי להגדיר מהו "דף איכותי". לדוגמא, נניח יוצרו של דף p באינטרנט כולל בו קישור לדף q , משמעות הדבר היא כי דף q חשוב לו במידה מסויימת ולכן הוא מהווה עבורו דף איכותי. יותר מכך, הקשרים נותנים לנו את ההזדמנות למצוא דפים איכותיים פוטנציאליים אך ורק על ידי הדפים המצביעים עליהם. העניין מאפשר לנו לעקוף את הבעיה שהועלתה מקודם שכן הרבה דפים אינם מספקים די אינפורמציה על עצמם. ישנם מספר חסרונות בשימוש בקישורים לצורך מטרה זו. א. קישורים באינטרנט נוצרו מסיבות שונות ולאו דווקא לצורך הקניית איכות לדף. לדוגמא, ישנם קישורים שמטרתם היחידה היא ניווט באינטרנט וקישורים לצורך פרסומות. ב. הקושי למצוא את האיזון המתאים בין רלוונטיות ופופולאריות. כאשר שני מושגים אלו תורמים לרעיון האיכות שאנו מדברים עליו. בעבודה זו, נציע מודל המבוסס על קשרים לצורך הקניית המושג דף איכותי ונראה כיצד הוא מוביל לשיטה המזהה דפים רלוונטיים איכותיים באינטרנט עבור שאילתא כלשהי. המודל מבוסס על הקשר הקיים בין מקורות הסמכות לנושא מסויים לבין אותם דפים המקושרים להרבה מקורות סמכות קרובים. נקרא לדפים מהסוג השני אלו-רכזים. נראה כי סוג מסויים של הקבלה טבעית זו קיימת בין רכזים לדפים איכותיים קיימת בגרף המוגדר על ידי מבנה הרשת המתוקשרת. ננצל זאת על מנת לפתח אלגוריתם שיזהה את הדפים משני הסוגים בו זמנית. האלגוריתם מופעל על תת גרף מרוכז של האינטרנט, אותו נבנה מהפלט שנקבל ממנוע החיפוש. בניית תת גרף זה תייצר אוסף קטן של דפים שיכילו ככל הנראה את הדפים הרלוונטיים ביותר עבור החיפוש.

**בניית גרף מרוכז באינטרנט** נוכל להסתכל על אוסף V של דפים מקושרים באינטרנט כעל גרף מכוון. G=(V,E) כך שהקודקודים יהיו הדפים וכל קשת מכוונת  מראה על קיומו של קישור מ p ל q. נגדיר את דרגת היציאה של קודקוד p להיות מספר הקודקודים שיש לו קישורים אליהם ואת דרגת הכניסה של קודקוד p להיות מספר הקודקודים שיש להם קישורים אליו. מתוך הגרף G נוכל לבודד אזורים קטנים או תתי גרפים באופן הבא: אם היא תת קבוצה של דפים, נסמן ב את הגרף המושרה מ .קודקודיו יהיו הדפים ב  והקשתות שלו יהיו כל הקישורים בין הדפים ב . נניח כי נתונה שאילתא בעלת נושא חיפוש רחב המצויינת בעזרת מחרוזת . נרצה לזהות את הדפים המוסמכים על ידי ניתוח מבנה הקישור. ראשית, אנו חייבים לזהות את תת הגרף באינטרנט עליו יפעל האלגוריתם. מטרתינו היא למקד את המאמץ החישובי בדפים הרלוונטים. לכן, לדוגמא, נוכל להגביל את הניתוח שלנו בקבוצה  של כל הדפים המכילים את המחרוזת . אך יש לכך שני חסרונות חשובים. ראשית, קבוצה זו יכולה להכיל מעל מיליון דפים ולכן תדרוש עלות חישובית רבה. שנית, קודם לכן ציינו כי חלק מהמקורות המוסמכים החשובים ביותר יכולים להיות לא מוכללים בקבוצה זו. באופן אידיאלי, נרצה להתמקד באוסף של דפים שהוא בעל התכונות הבאות: א. קטנה יחסית. ב.  עשירה בדפים רלוונטיים. ג.  מכילה את רוב ( או הרבה) ממקורות הסמכות החזקים. על ידי כך שנשמור אתלהיות קטנה נוכל להרשות לעצמינו את העלות החישובית על ידי שימוש באלגוריתמים לא טריוויאליים. על ידי הבטחת העושר בדפים הרלוונטיים אנו הופכים את התהליך למציאת מקורות סמכות טובים לקל יותר מכיוון שסביר להניח כי יש אליהם הפנויות רבות בתוך . כיצד נוכל למצוא אוסף כזה של דפים? עבור פרמטר t (בדרך כלל הוא בערך 200),קודם לכן ניקח את t הדפים המדורגים ביותר עבור המחרוזת  על ידי מנוע חיפוש כגון AltaVistaאו Hotbot. מנועי חיפוש אלו היו קיימים באותה התקופה, לפני קיומו של Google. נתייחס ל t דפים אלו כאל קבוצה שורשית . קבוצה זו מקיימת את תנאים א' ו ב' שצויינו קודם אך היא רחוקה מאוד מלקיים את תנאי ג'. על מנת לראות זאת, נשים לב כי t הדפים שבראש הרשימה המוחזרים על ידי מנוע החיפוש כולם יכללו את המחרוזת ולכן גם את . קבוצה זו היא בוודאות תת קבוצה של כל האוסף  של כל הדפים המכילים את . עם זאת טענו כי אפילו  בדרך כלל אינה מקיימת את תנאי ג'. מעניין לציין כי בדרך כלל יש מעט מאוד קישורים בין דפים ב , מה שמשאיר אותה למעשה חסרת צורה. לדוגמא, בניסויים שלנו, הקבוצה השורשית עבור השאילתא "java" מורכבת מ 15 קישורים בין דפים מתחומים שונים. נוכל להשתמש בקבוצה השורשית על מנת לייצר את קבוצת הדפים שתענה על הדרישות שלנו. אם ניקח מקור סמכות חזק עבור שאילתת חיפוש רחב, ייתכן והוא לא יהיה כלול בקבוצה , אך סביר להניח שלפחות אחד הדפים ב  יצביע עליו על ידי קישור. לכן נוכל להגדיל את מספר מקורות הסמכות החזקים על ידי הרחבת הקבוצה . בצורה זו נשיג את הקבוצה  המתקבלת על ידי הרחבת הקבוצה בעזרת הוספת כל הדפים שישנם קישורים ב המצביעים עליהם. על מנת להבין זאת נסתכל על הציור הבא:

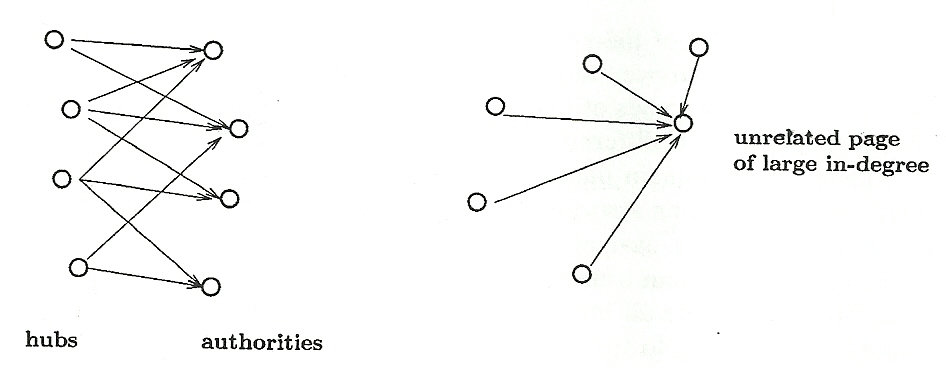




נרצה לעשות הגבלה על ידי כך שנאפשר לכל דף ב  להביא לכל היותר d דפים המצביעים על דפים שיהפכו ל . פרט זה הוא חשוב מכיוון שאם נוסיף לקבוצה דפים ללא הגבלה נפגע בתכונה א' הדורשת שהקבוצה  תהיה קטנה יחסית.

ישנה שיטה שהיא שימושית מאוד לצורך קיזוז השפעת הקישורים שמטרתם היחידה היא ניווט בתוך האינטרנט. נניח ישנו  כמו שתיארנו קודם. נבדיל בין שני סוגים של קישורים בתוך . קישור חיצוני, הוא קישור המקשר בין דפים עם שמות תחום שונים וקישור פנימי, המקשר בין דפים עם אותו שם תחום. קישור חיצוני הוא קישור היוצא ומצביע על דף של ארגון אחר. קישור פנימי הוא יוצא ומצביע על דף בתוך הארגון עצמו. הכוונה ב"שם תחום" כאן היא הרמה הראשונה של מחרוזת ה URL הקשורה לעמוד (Domain). כלומר, החלק הראשון של כתובת האינטרנט של הדף. בדרך כלל, מטרתם של הקישורים הפנימיים היא ניווט בתוך התשתית של האתר ולכן הם מספקים הרבה פחות אינפורמציה על הדפים האיכותיים שהם מצביעים אליהם. לכן נמחק את כל הקישורים הפנימיים של הגרף  ונשאיר רק את הקשתות המתייחסות לקישורים החיצוניים. התוצאה תהיה הגרף .

**חישוב רכזים ודפים איכותיים** ראינו כיצד ניתן ליצור תת גרף קטן  שהוא מתמקד יחסית בנושא השאילתא. יש בו הרבה דפים רלוונטיים ומקורות סמכות חזקים. עכשיו נראה כיצד לחלץ את מקורות הסמכות מתוך אוסף כל הדפים. נעשה זאת אך ורק על ידי ניתוח הקשרים של . ניתן לטעון כי השיטה הפשוטה ביותר תהיה לסדר את הדפים על ידי דרגת הכניסה שלהם-מספר הקישורים אשר מצביעים עליהם ב . קודם לכן, דחינו את הרעיון כי היה מדובר בכל הדפים המכילים את המחרוזת . כעת יש לנו אוסף קטן של דפים רלוונטיים שבנינו במפורש, המכיל בתוכו את כל מקורות הסמכות שאנו רוצים למצוא. לכן מקורות סמכות אלו הם גם שייכים ל  וגם יש להם הפניות בצורה רבה על ידי דפים מתוך . אכן, גישת הדרוג המתבצעת אך ורק על ידי דרוג דרגת הכניסה עובדת טוב יותר בהקשר של  מאשר במסגרת שהצענו קודם לכן. במקרים מסויימים, זה יכול לייצר באופן אחיד תוצאות ברמה גבוהה. בכל זאת, לשיטה עדיין כמה בעיות רציניות. לדוגמא, עבור השאילתא "java" הדפים בעלי דרגת הכניסה הגבוהה ביותר מורכבים מ [www.gamelan.com](http://www.gamelan.com) ו java.sun.com יחד עם דפי פרסומות לחופשות באיים הקאריביים ודף הבית של Amazon Books. תערובת זו מציגה סוג של בעיה שיכולה לצוץ בעקבות שיטה של דרוג פשוט. בזמן ששני הדפים הראשונים יכולים להיחשב כתוצאות טובות, השאר אינם רלוונטיות לנושא השאילתא. יש להם דרוג גבוה אך הם חסרות אחדות נושאית. הקושי הבסיסי שנגרם מכך הוא המתח הטמון בגרף  בין מקורות הסמכות החזקים לבין דפים שהם פשוט פופולאריים בצורה אוניברסאלית. אנו מצפים מהסוג השני של הדפים דרגת כניסה גבוהה ללא תלות בנושא השאילתא הבסיסי. ניתן לתהות האם עקיפת בעיות אלו לא תדרוש לעשות שימוש עתידי בתוכן המילולי של של קבוצת הבסיס יותר מאשר במבנה הקישורי של . כעת נראה שזה לא המקרה. כלומר למעשה אפשרי לדלות אינפורמציה בצורה אפקטיבית יותר מהקשרים מאשר מהתוכן המילולי. נתחיל מהעובדה הבאה. דפים סמכותיים הרלוונטיים לשאילתא ההתחלתית, לא רק שהם אמורים להיות עם דרגת כניסה גבוהה, אלא מכיוון שהם כולם מקורות סמכות בעלי נושא משותף, אמורה להיות גם חפיפה רבה בקבוצות הדפים המצביעים עליהם. לכן, בנוסף לדפים בעלי הסמכות הגבוהה נצפה למצוא דפים רכזים. דפים רכזים אלו דפים שיש בהם קישורים לדפים איכותיים רבים. דפים רכזים אלו מקבצים יחד דפים איכותיים בעלי נושא משותף ומאפשרים לנו לזרוק את הדפים הלא רלוונטיים בעלי הדרוג הגבוה. נסתכל על הציור הבא לצורך המחשה.



דפים רכזים ודפים איכותיים מקיימים יחס של חיזוק הדדי. דף רכז טוב הוא דף המצביע על הרבה דפים איכותיים. דף איכותי טוב הוא דף שמצביעים עליו הרבה דפים רכזים טובים. ניתן להגיד בברור כי על מנת לזהות דפים רכזים ודפים איכותיים של תת גרף  אנו צריכים שיטה שתשבור מעגליות זו.

**האלגוריתם האיטרטיבי**

נשתמש בקשר בין הדפים הרכזים לדפים האיכותיים בעזרת אלגוריתם איטרטיבי המתחזק ומעדכן את המשקלים המספריים של כל דף. בהינתן גרף מכוון, נחלק משקלים לקודקודים. לדף p נתאים משקל סמכות  ומשקל רכזות . נרצה שהמשקלים יהיו מנורמלים כלומר , . כעת, נבצע את התהליך האיטראטיבי הבא:תחילה נחלק משקלי רכזות לקודקודים כלומר את . למצב זה נקרא . כעת, נבצע לפי התור את הפונקציות הבאות:

, .

כלומר, נתחיל מוקטור משקלי רכזות אקראי מנורמל ואז סדר המשקלים יהיה:



למעשה, אם מטריצת הסמיכויות של הגרף היא  אז הפעולות הן:

, .

נסביר זאת באמצעות דוגמא:

נסתכל בגרף המכוון הבא, נניח כי הקודקודים הם:p1, p2, p3.

Y = משקל הרכזות. הסכום של אילו שהוא מצביע עליהם.

X = משקל האיכות. הסכום של אילו שמצביעים עליו.

נניח כי משקלי הרכזות הם:a, b, c)).

b

a

c

במקרה זה, מטריצת הסמיכויות היא: והמטריצה המשחולפת היא . נשים לב כי מטריצת הסמיכויות היא לא סימטרית מכיוון שהגרף הוא מכוון.

טענה: עבור כל מטריצות ו  הן סימטריות (אך לאו דווקא שוות).

אם נעדכן את משקלי הסמכות נקבל את הוקטור (c, a, a+b) וזה בדיוק:

. כנ"ל אם וקטור משקלי הסמכות היה a, b, c)) והיינו רוצים לעדכן את וקטור משקלי הרכזות אז באמת היינו מקבלים  כנדרש.

לפי הגדרת הפעולות שדיברנו עליהם:  כעת,

.

מכאן,



כמו כן,



הערה: למעשה בכל שלב אנו מנרמלים את הוקטור המתקבל, אם כי לא ציינו זאת במפורש.

נסמן: . סימטרית לפי הטענה. קבלנו כי : .

כל התהליך שעשינו מקביל Power Method רק שהוא מתבצע בשני שלבים. לכן לפי ה Power Method נקבל כי מתכנס לוקטור העצמי  של המתאים לערך העצמי הגדול ביותר של .

השאלה היא למה זו בדיוק נקודת שבת?

התהליך מורכב משני שלבים:

אם נמצא זוג כך ש: וגם , נוכל לנרמל על ידי חלוקה ב  ו  בהתאמה ואז נקבל כמעין נקודת שבת כי:  כלומר ו . לכן היא נקודת השבת של המטריצה .

המסקנה היא כי כל התהליך מקביל ל Power Method.

כמו כן ניתן להגדיר מטריצה  ובאותה הדרך להגיע למצב ואז  כלומר , .

הגענו לתוצאה כי ל B ו C אותם ערכים עצמיים אך וקטורים עצמיים שונים.

מסקנה: בעזרת תהליך ה Power Method הגענו לוקטור המשקלים המתאים לערך העצמי המקסימלי. וקטור זה מהווה נקודת שבת ולכן נקבל דרוג לקודקודים.

הערה: האלגוריתם שהוצג כאן הוא רק אחד מתוך כמה שיקולים בעזרתם מנועי החיפוש מבצעים את הדרוג. הדרוג הסופי נבנה בעזרת איזון בין כמה שיקולים. לפי התוצאות המוצגות במאמר ההתכנסות של Iterate היא די מהירה. למעשה, הממצאים הם כי מספיקות כ-20 אינטראציות כדי להתכנס להגיע למצב שבו כמעט ולא זזים מנקודת השבת. אחרי שהתכנסנו לנקודת השבת ניתן להגדיר את המספר . c מספר זה ייצג את כמות הדפים הראשונים בדרוג שנרצה שיוצגו על ידי מנוע החיפוש.

רשימה ביבליוגרפית

•JON M. KEINBERG/Authoritative Sources in a Hyperlinked Environment/*Journal of the ACM, Vol. 46, No. 5, September 1999, pp. 604-632.*

FAN R. K. CHUNG/ Spectral Graph Theory, 1994. •

• Chris Godsil and Gordon Royle/ *Algebraic Graph Theory,2001.*

•Lecture site of Prof. Dan Spielman.

<http://www.cs.yale.edu/homes/spielman/561/>

•Website of Google.

<http://www.googleguide.com/google_works.html>